ABSTRACT OF THE DISCLOSURE

APPARATUS AND METHODS FOR MAXIMIZING SERVICE-LEVEL-AGREEMENT PROFITS

Apparatus and methods for maximizing service-level-agreement (SLA) profits are provided. The apparatus and methods consist of formulating SLA profit maximization as a network flow model with a separable set of concave cost functions at the servers of a Web server farm. The SLA classes are taken into account with regard to constraints and cost function where the delay constraints are specified as the tails of the corresponding response-time distributions. This formulation simultaneously yields both optimal load balancing and server scheduling parameters under two classes of server scheduling policies, Generalized Processor Sharing (GPS) and Preemptive Priority Scheduling (PPS). For the GPS case, a pair of optimization problems are iteratively solved in order to find the optimal parameters that assign traffic to servers and server capacity to classes of requests. For the PPS case, the optimization problems are iteratively solved for each of the priority classes, and an optimal priority hierarchy is obtained.